# Submission in Response to NSF CI 2030 Request for Information

**DATE AND TIME:** 2017-04-05 16:31:54
**REFERENCE NO:** 289

### Author Names & Affiliations

- Richard Signell - USGS, US Integrated Ocean Observing System (IOOS), Earth System Information Partners (Interoperability Committee Chair)
- Emilio Mayorga - University of Washington, US Integrated Ocean Observing System (US-IOOS)

### Contact Email Address (for NSF use only)

(Hidden)

### Research Domain, discipline, and sub-discipline

Physical Oceanography, Chemical Oceanography, River Biogeochemistry

### Title of Submission

Going from "Working to Working": Supporting and leveraging existing successful frameworks for NSF benefit

### Abstract (maximum ~200 words).

We believe a major challenge for NSF is how to best leverage and support the thriving, functioning standards-based frameworks for data discovery and access already in widespread community use. NSF could play an important role in training researchers to use these frameworks, strategically expanding them to maximize benefit for the NSF research community, and making them more robust and scalable. Successful community-driven cyberinstructure development like the NSF-supported Unidata Program Center, and emerging EarthCube and DataOne contributions that become adopted should be copied and expanded. In addition, a mechanism to fund small projects that could address open issues on key framework components could dramatically accelerate progress, yielding large, reliable returns on investment.

### Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

How to best leverage and support the successful ecosystem of data discovery and access tools that have emerged to maximize benefit for the NSF research community?

NSF cyberinfrastructure research seeks to foster the development of a scalable, comprehensive, secure and sustainable cyberinfrastructure that supports potentially transformative research in science and engineering. Part of this effort is to foster technologies that make it possible for NSF researchers to more efficiently and effectively discover and access data, which in addition to allowing potentially transformative research, simply reduces the time spent on routine data wrangling tasks, leaving more time for actual science.

Fortunately, decades of investment by the community have resulted in a thriving ecosystem of tools that enable standardized search and access for a variety of scientific data types across a widely-distributed population of data providers. For providers, it is now easy to deliver earth data via standardized web services and transform existing information about the data into standardized metadata. For consumers, it is now easy to discover and access earth data from point collections, time series, profiles, trajectories, and gridded data in a common form, allowing interoperable data access across institutional and discipline boundaries.

The NSF-funded Unidata Program Office has played an important role in the development of the ecosystem, delivering key infrastructure components such as NetCDF, OPeNDAP, the THREDDS Data Server, and common representation of scientific feature types. In contrast to being driven by principal investigator interests, Unidata is driven by user community needs, and operates on the Jim Gray law of going from “Working to Working”, continually improving and building new components that work within the existing ecosystem. They also provide support and training for their tools. Such engagement and support benefit the wider community and should be sustained. What began as an effort to develop tools for atmospheric modeling now supports modeled and observed data from meteorology, oceanography and hydrology, and is poised for future expansion. EarthCube and DataOne have also contributed some useful components to this ecosystem.

While NSF has played an important role, much of the ecosystem has been developed by others in the community:

- The Open Geospatial Consortium (OGC) created geospatial standards such as Catalog Services for the Web (CSW), Web Map Service (WMS) and Sensor Observation Services (SOS).

- NOAA built important components that transform data in NetCDF files or OPeNDAP datasets into ISO metadata, and developed the ERDDAP data server, widely used for it’s RESTful interface to regular gridded and tabular data.

- NASA funded the Earth System Information Partners (ESIP) to develop Attribute Conventions for Data Discovery (ACDD) and HDF5 technologies used in NetCDF4.

- The US Integrated Ocean Observing System (IOOS) added SOS capabilities to the THREDDS Data Server.

- The British Met Office built a Python package that supports interoperability enabled by the CF conventions.

- The commercial company Mapbox supports development of the open-source Leaflet toolkit for displaying 2D maps in the browser.

- The commercial company AGI supports development of the open-source Cesium toolkit for displaying 3D maps in the browser.

- The Australian National Government built the open-source TerriaJS web portal framework that uses Leaflet and Cesium to display data from standard web map services.

- The Sloan Foundation funded the Jupyter Project, which allows for open, reproducible and exploratory data science in the browser.

Thus a major challenge for NSF is how to best leverage and support this ecosystem of data discovery and access tools to maximize benefit for the NSF research community. While the progress on this standards-based ecosystem is exciting, there is still much work to be done, including expanding the use of these tools to other disciplines, making them work in all the dominant environments that scientists use, and making the tools more capable, scalable and robust.

**Question 2** Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Some type of mechanism to fund small projects or even some mechanism by which funding could be directed to certain specific tasks (with some criteria for selection and oversight) would be extremely useful (e.g. along the lines of EAGER awards). Here are some examples where small amounts of investment would yield large returns:

NetCDF-Java and THREDDS Data Server. The next generation models in atmosphere, groundwater and ocean modeling all use grids that are not logically rectangular (e.g. triangular, hex mesh). The community has developed a standard conventions for representing these grids, but Unidata doesn’t have the manpower to add support to their netcdf-java library, so the new forecast models are not supported. Unidata could provide support if they could hire just one extra person, at a cost of approximately $200K/year.

ERDDAP. ERDDAP is an very useful tool for transforming collections of uniform grids and tabular geoscience data into standardized services that utilize a RESTful interface, allowing interfacing to Python, Matlab, R, web applications and more. It is deployed at more than than 50 locations including NOAA, USGS, every region of the US Integrated Ocean Observing System and in more than 10 countries around the world. Yet it is maintained by a single NOAA employee, which is both encouraging (in that one employee can develop, maintain and support it) and worrisome (it would be nice to have a broader development base for stability). Again, Unidata could provide significant support if they could hire just one extra person, at a cost of approximately $200K/year.

pycsw. pycsw is used to provide OGC Catalog Service for the Web (CSW) support for data.gov, data.gov.uk, data.ioos.us and more. It has 54 open issues, and is hoping to get a Google Summer of Code student to build a Solr backend (fast, open source enterprise search platform built on Apache Lucene). Small amount of funding could just make this happen.

Matlab toolbox for access to standardized web services. Currently only Python users have tools to query standard catalog services (CSW, OpenSearch), and enjoy interoperable access to CF-DSG and CF-UGRID scientific feature types (profiles, trajectories, structured grids, unstructured grid models), but there are no tools for Matlab users. One might hope that Mathworks would support such a toolbox, but the geoscience community is too small. Yet Matlab is still the primary tool used by certain NSF communities (like oceanography and many engineering disciplines) so NSF could help more of their community to be able to efficiently access data by supporting development of a Matlab toolbox to access standard services.

TerriaJS. This powerful open source framework for web portal development is under heavy development and has 250 issues open. Targeting specific issues important for specific NSF communities, perhaps through something like BountySource.

Jupyter Notebooks. Developing examples that illustrate end-to-end workflows could be hugely useful for teaching other researchers how to use web services to search and discover data. Something along the lines of Unidata’s Notebook Gallery for example.

Ocean Observatories Initiative Cyberinfrastructure (OOI-CI). The NSF funded OOI-CI has built their own non-standard proprietary infrastructure and thus far has not produced any tools that are in use by the community. Yet if they made a small investment in adopting the standardized approach used elsewhere in the community, they could dramatically cut their operating costs, improve their ability to serve data to the research community, and contribute to the issues above.

### Consent Statement